Penggunaan Algoritma *Breadth-First Search* Pada *Web Crawler* Untuk Memblokir Situs Pornografi Anak

Gianfranco Fertino Hwandiano / 13515118
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung
13515118@std.stei.itb.ac.id

***Abstrak*—Teknologi *Web Crawler* memungkinkan pihak berwenang untuk mendeteksi situs-situs yang mengandung pornografi anak yang sudah sangat jelas dilarang oleh undang-undang. Dengan teknologi ini, pencarian situs-situs tersebut bisa dilakukan dengan mudah tanpa harus mencari satu persatu secara *manual* (cara konvensional). Teknologi *Web Crawler* ini memanfaatkan algoritma *Breadth First Search,* yaitu algoritma pencarian graf traversal.**

Kata Kunci—webcrawler, pornografi anak, bfs, graf.

I. PENDAHULUAN

Pornografi anak biasanya diperoleh oleh pedofil yang menggunakan gambar untuk berbagai keperluan, mulai dari menggunakannya untuk kepentingan seksual pribadi, perdagangan dengan pedofil lain, menyiapkan anak-anak untuk pelecehan seksual sebagai bagian dari proses yang dikenal sebagai "perawatan anak", atau bujukan yang mengarah ke jebakan untuk eksploitasi seksual seperti produksi pornografi anak yang baru atau prostitusi anak.[1]

Gambar 1. *Stop Child Porn*

Di Indonesia sendiri, pornografi terutama pornografi anak sangatlah dilarang. Hal tersebut diatur dalam Undang-Undang pasal 4 ayat (1) UU 44/2008 yang mengatur larangan perbuatan memproduksi, membuat, memperbanyak, menggandakan, menyebarluaskan, menyiarkan, mengimpor, mengekspor, menawarkan, memperjualbelikan, menyewakan, atau menyediakan pornografi yang secara eksplisit memuat salah satunya adalah pornografi anak.[2]

Sudah banyak sekali kasus pornografi yang melibatkan anak, termasuk di Indonesia. Berita terbaru berikaitan dengan hal ini di Indonesia adalah kasus pornografi anak yang dilakukan melalui media sosial. Nama grup dari komplotan ini adalah "Official Candys Group". Grup yang dibentuk September 2016 ini telah beranggotakan 7.479 *members*.. Sudah ada delapan anak perempuan yang menjadi korban pelecehan seksual yang telah diketahui sampai saat ini.[3] Ini merupakan masalah serius yang tidak bisa diabaikan.

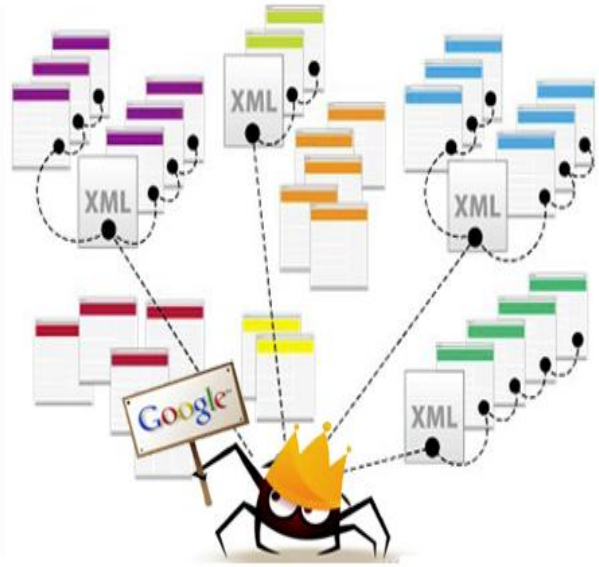
Untuk memblokir situs-situs yang mengandung pornografi anak, digunakanlah teknologi *Web Crawler*. Teknologi ini diimplementasi menggunakan algoritma *Breadth-First Search*. Dengan teknologi ini, situs-situs terlarang tersebut dapat terdeteksi secara otomatis.

II. DASAR TEORI

A. Web Crawler

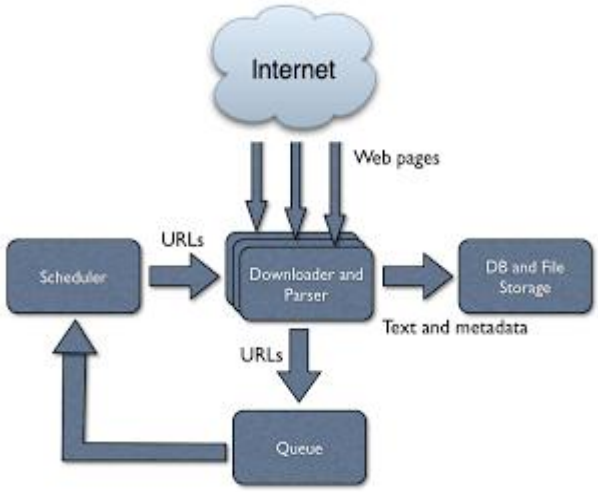
Web crawler adalah sebuah perangkat lunak yang digunakan untuk menjelajah serta mengumpulkan halaman-halaman web yang selanjutnya diindeks oleh mesin pencari. Desain sebuah *web crawler* harus dapat memenuhi berbagai tujuan yang kadang saling berbenturan, seperti kualitas halaman yang diambil, penyebaran dan pengurangan beban jaringan, kecepatan pengambilan, serta pada saat yang bersamaan tidak membebani server situs yang dikunjungi. *Web crawler* juga sering dikenal dengan nama *web spider atau web robot*, yaitu salah satu komponen penting dalam arsitektur sebuah mesin pencari modern. Fungsi utama *web crawler* adalah untuk melakukan penjelajahan dan pengambilan halaman-halaman web yang ada di internet. Hasil pengumpulan situs web selanjutnya akan diindeks oleh mesin

pencari sehingga mempermudah pencarian informasi di internet.



Gambar 2. Illustrasi dari *Web Crawler*

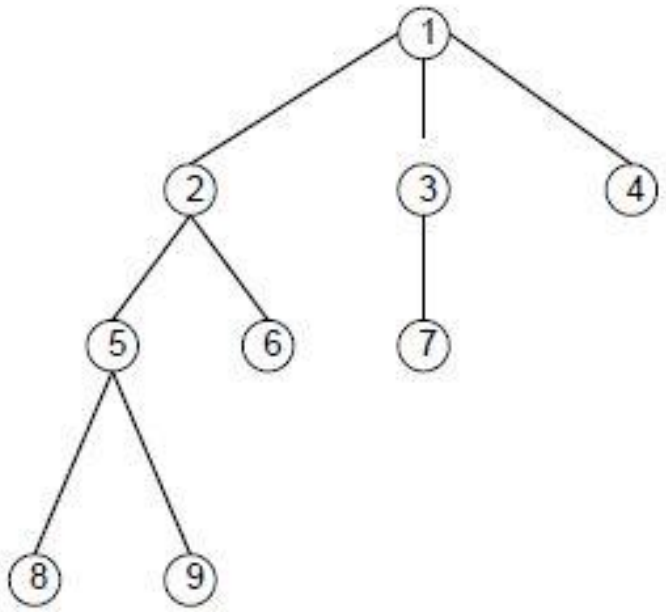
Dalam kerjanya, *web crawler* melakukan proses validasi terhadap HTML *code* dan tautan yang ditemukan pada situs yang dikunjungi. Setelah melakukan validasi, halaman web dan tautan tersebut di*download* kemudian diparsing ke dalam tempat penyimpanan utama. *Data* yang dibawa oleh *web crawler* secara sederhana hanya berupa *text* dan metadata. Sedangkan data tautan yang ditemukan pada halaman *web* yang dikunjungi akan ditempatkan pada seeds (tempat penyimpanan data URL) masuk ke dalam antrian kunjungan selanjutnya *web crawler*. Secara simultan *web crawler* mengunjungi situs-situs yang alamatnya terdapat di dalam antrian sampai data URL habis atau dihentikan oleh administrator.[4]



Gambar 3. Arsitektur sebuah *Web Crawler*

B. Algoritma Breadth-First Search

Pencarian melebar (*Breadth First Search* atau BFS) merupakan pencarian traversal graf yang melakukan pencarian secara melebar sesuai namanya. Misalkan kita mempunyai graf G yang mempunyai n buah simpul. Kita akan melakukan traversal di dalam graf dan misalkan traversal dimulai dari simpul v. Algoritma BFS adalah sebagai berikut : kunjungi simpul v, kemudian semua simpul yang bertetangga dengan simpul v dikunjungi terlebih dahulu. Selanjutnya, simpul yang belum dikunjungi dan bertetangga dengan simpul-simpul tadi dikunjungi, demikian seterusnya. Jika graf berbentuk pohon beakar, maka semua simpul pada aras *d* dikunjungi lebih dahulu sebelum simpul-simpul pada aras *d*+1.



Gambar 4. Contoh Graf

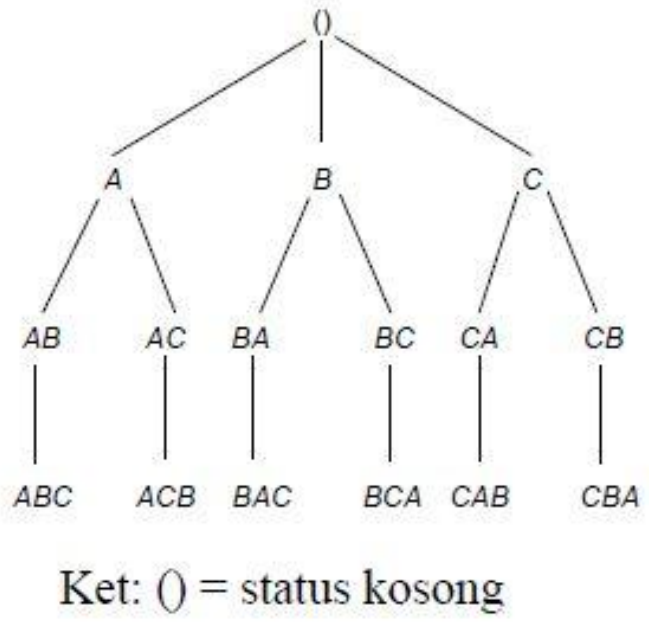
Tinjau graf pada gambar 3. Bila graf dikunjungi mulai dari simpul 1, maka urutan simpul yang dikunjungi adalah 1,2,3,4,5,6,7,8,9.

Algoritma BFS memerlukan sebuah antrian *q* untuk menyimpan simpul yang telah dikunjungi. Simpul-simpul yang telah dikunjungi suatu saat diperlukan sebagai acuan untuk mengunjungi simpul-simpul yang bertetangga dengannya. Tiap simpul yang telah dikunjungi masuk ke dalam antrian hanya satu kali.

Berikut adalah *pseudocode* dari BFS :

```
procedure BFS(input v:integer)
{ Traversal graf dengan algoritma pencarian
BFS.
Masukan: v adalah simpul awal kunjungan
Keluaran: semua simpul yang dikunjungi dicetak
ke layar
}

Deklarasi

      w : integer
      q : antrian

procedure BuatAntrian(input/output q : antrian)
{ membuat antrian kosong, kepala(q) diisi 0 }


procedure MasukAntrian(input/output q:antrian,
input v:integer)
{ memasukkan v ke dalam antrian q pada posisi
belakang }

procedure HapusAntrian(input/output
q:antrian,output v:integer)
{ menghapus v dari kepala antrian q }

function AntrianKosong(input q:antrian)
<- boolean
{ true jika antrian q kosong, false jika
sebaliknya }

Algoritma:

BuatAntrian(q) { buat antrian kosong }

write(v) { cetak simpul awal yang dikunjungi }
dikunjungi[v] <- true { simpul v telah
dikunjungi, tandai dengan
true}
MasukAntrian(q,v) { masukkan simpul awal
kunjungan ke dalam
antrian}

{ kunjungi semua simpul graf selama antrian
belum kosong }
while not AntrianKosong(q) do
      HapusAntrian(q,v) { simpul v telah
dikunjungi, hapus dari
antrian }
      for tiap simpul w yang bertetangga
      dengan simpul v do
            if not dikunjungi[w] then
                  write(w) {cetak simpul yang
                  dikunjungi}
                  MasukAntrian(q,w)
                  dikunjungi[w] <- true
            endif
      endfor
endwhile {AntrianKosong(q)}
```

Cara untuk memvisualisasikan proses dari BFS adalah dengan menggunakan pohon ruang status. Simpul-simpul di dalam pohon dinamis yang memenuhi kendala menyatakan **status persoalan** (*problem state).* Suatu operator mentransformasikan persoalan dari sebuah status ke status yang lain. Solusi persoalan dinyatakan dengan satu atau lebih status yang disebut **status solusi** (*solution state).* Status solusi yang merupakan simpul daun disebut **status tujuan** (*goal state).* Himpunan semua status solusi disebut **ruang status (***state space)* dan pohonnya dinamakan juga **pohon ruang status** (*state space tree).* Akar pada pohon ruang status menyatakan **status awal** (*initial state)*sedangkan daun menyatakan status solusi.[6]



Gambar 5. Pohon ruang status pembangkitan permutasi A,B,C

III. IMPLEMENTASI

A. Cari Sumber Pornografi Anak

Langkah awal yang dilakukan adalah mencari sumber pornografi anak. Umumnya pihak berwenang seperti polisi menyimpan benda seperti ini sebagai berkas bukti atas kasus yang pernah ditanganinya. Dalam makalah ini, karena penulis tidak mempunyai akses terhadap sumber yang berkaitan, maka penulis menggunakan alamat situs biasa yang tidak mengandung pornografi. Makalah ini dibuat hanya sebagai bahan edukasi.

Sumber berupa daftar alamat situs yang telah tertangkap basah sebelumnya karena mengandung pornografi anak. Setiap alamat situs yang kita miliki ini nantinya kita terapkan mesin *web crawler*. Selain itu, proses *crawling* biasanya dilakukan di *deep web* yaitu bagian lain dari internet yang membutuhkan akses khusus.

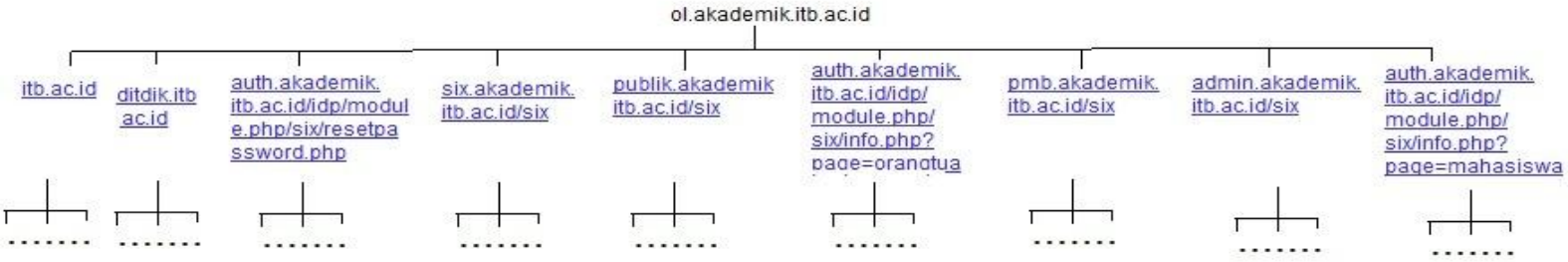
B. Membuat Web Crawler

Cara kerja dari *web crawler* dengan algoritma *Breadth First Search* adalah sebagai berikut :

1. Kita buat dua antrian. Antrian pertama (kita sebut q1) fungsinya untuk mensimulasikan proses *Breadth First Search* (proses *enqueue* dan *dequeue* terjadi disini). Antrian kedua fungsinya untuk menampung semua alamat situs yang pernah di-*crawl*. Antrian inilah nantinya yang berperan sebagai *search result* yang berisi alamat-alamat situs yang mungkin berhubungan atau memuat konten situs pornografi anak.
2. Kita pilih sebuah URL sebagai akar dari pencarian kita.
3. Kemudian *web crawler* akan mengambil HTML (*HyperText Markup Language*) kode dari URL tersebut.
4. HTML tersebut kemudian di-*parsing* isinya untuk kita ambil semua URL yang terdapat dalam situs tersebut. Hal ini bisa dilakukan dengan mencari tag URL dari kode HTML yang didapat.
5. Setiap sebuah URL baru ditemukan, kita cek dahulu apakah kita sudah melakukan crawl terhadap situs tersebut. Jika sudah maka abaikan (jangan *push).* Hal ini dilakukan untuk mencegah terjadinya *infinite process* pada proses *crawling.* Jika belum maka alamat situs ini langsung di-*enqueue* ke dalam antrian pertama dan antrian kedua sesuai dengan prinsip *Breadth First Search*.
6. Setelah parsing selesai, kita ambil elemen terdepan dari antrian pertama (q1) atau yang biasa disebut *dequeue*.
7. Elemen yang terambil tadi yang berupa URL kemudian dijadikan akar pencarian berikutnya.
8. Ulangi terus proses ini hingga tidak ada lagi elemen yang tersisa dalam antrian.
9. Setelah selesai, output semua situs yang pernah di-*crawl* yang tersimpan dalam antrian kedua ke dalam file .txt untuk nantinya diselidiki lebih lanjut.

Berikut adalah potongan kode dari *web crawler* yang dibuat dengan bahasa Java

```
import org.jsoup.Jsoup;
import org.jsoup.nodes.Document;
import org.jsoup.nodes.Element;
import org.jsoup.select.Elements;
import java.io.IOException;

public class BasicWebCrawler {

    public static void main(String[] args) {
        // initial web page
        String s = args[0];
        // list of web pages to be
        examined
        Queue<String> queue = new
        Queue<String>();
        queue.enqueue(s);
        // set of examined web pages
        SET<String> marked = new
        SET<String>();
        marked.add(s);
        // breadth first search crawl of
        web
        while (!queue.isEmpty()) {
            String v = queue.dequeue();
            StdOut.println(v);
            //2. Fetch the HTML code
            Document document =
            Jsoup.connect(URL).get();
            //3. Parse the HTML to extract
            links to other URLs
            Elements linksOnPage =
            document.select("a[href]");

             //5. For each extracted URL..
            for (Element page : linksOnPage)
            {
                queue.enqueue(page.attr("abs:
                href"));
                marker.add(page.attr("abs:hre
                f"));
            }
        }
    }
}
```

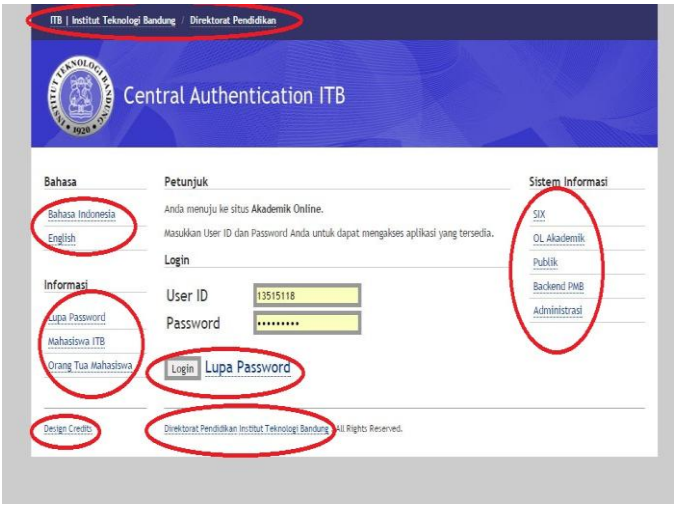


Gambar 7. Pohon Ruang Status

C. Hasil dan Analisis

Dalam makalah ini, karena penulis tidak mempunyai akses terhadap sumber yang berkaitan (situs yang mengandung konten pornografi anak) , maka penulis menggunakan alamat situs biasa yang tidak mengandung pornografi._Sebagai studi kasus, penulis memilih www.ol.akademik.itb.ac.id sebagai bahan uji.

Web Crawler kemudian melakukan *Http Request* terhadap ol.akademik.itb.ac.id. Setelah itu, diambilah *webpage* tersebut dalam bentuk HTML. Kode HTML yang diambil tersebut kemudian di-*parsing* untuk didapatkan URL yang terkandung.



Gambar 6. Situs ol.akademik.itb.ac.id

Tinjau gambar 6 tersebut. Dari situs ol.akademik.itb.ac.id, kita dapat mendapatkan sejumlah alamat situs yang dapat kita *crawl*, ditandai dengan lingkaran merah. Alamat-alamat situs tersebutlah yang nantinya akan dimasukan ke dalam antrian.

Setelah melewati proses pengambil kode HTML berikut adalah isi dari antrian setelah *crawling* pertama dilakukan :

http://www.itb.ac.id
http://www.ditdik.itb.ac.id
https://auth.akademik.itb.ac.id/idp/module.php/six/resetpassword.php
https://six.akademik.itb.ac.id/six/
https://publik.akademik.itb.ac.id/seven/
https://pmb.akademik.itb.ac.id/six/
https://admin.akademik.itb.ac.id/six/
https://auth.akademik.itb.ac.id/idp/module.php/six/info.php?page=mahasiswa
https://auth.akademik.itb.ac.id/idp/module.php/six/info.php?page=orangtua

Kemudian setelah itu, diambil elemen pertama dari antrian yaitu http://www.itb.ac.id. Situs tersebut kemudian diambil kode HTML nya lalu diparsing untuk didapatkan alamat-alamat situs yang terkandung di dalamnya.

Demikian seterusnya hingga semua situs di-*crawl*. Proses *crawling* ini memang memakan waktu yang cukup lama. Kecepatannya sangat bergantung pada kecepatan internet dari pengguna *web crawler*. Hal ini dikarenakan setiap situs diproses, kita melakukan *http request* terhadap situs tersebut dan mengunduh kode html dari situs tersebut yang ukurannya beragam.

Crawling terhadap situs ol.akademik.itb.ac.id ini menghasilkan banyak situs yang tidak mungkin dimuat semua disini.

Pohon ruang status dari permasalahan ini dapat dilihat pada gambar 7. Hanya ditampilkan sebagian karena tidak akan muat.

Dari kumpulan situs yang didapat dari hasil *crawling*, kita dapat melakukan penyelidikan lebih lanjut. Umumnya hampir semua situs yang memiliki hubungan dengan situs akar yang dijadikan pusat pencarian (yang telah terbukti mengandung konten pornografi anak) juga mengandung konten terlarang tersebut. Kedepannya tinggal bagaimana pihak yang berwenang bertindak terhadap situs-situs terlarang ini.

IV. KESIMPULAN

Pengaplikasian algoritma *Breadth First Search* dalam *Web Crawler* untuk pencarian situs yang mengandung konten pornografi sangatlah penting. Dengan adanya *Web Crawler*, pencarian situs terlarang tersebut menjadi lebih mudah karena prosesnya terautomatisasi. Kasus ini tidak bisa disepelekan karena taruhannya adalah masa depan anak muda penerus bangsa.

V. UCAPAN TERIMA KASIH

Rasa terima kasih penulis ucapkan terhadap Tuhan Yang Maha Esa karena hanya berkat rahma dan karunia-Nya makalah ini dapat diselesaikan. Penulis juga mengucapkan terima kasih kepada dosen pengajar mata kuliah IF2211 Strategi Algoritma, yaitu Dr. Masayau Leylia Khodra S.T., M.T., Dr. Ir. Rinaldi Munir, dan MT, Dr. Nur Ulfa Maulidevi, S.T., M.Sc. yang telah memberikan dasar ilmu untuk Strategi Algoritma.

DAFTAR PUSTAKA

[1] Lanning, Kenneth (2010). "Child Molesters: A Behavioral Analysis" (PDF). National Center for Missing & Exploited Children.

[2] Hukumonline.com, Sanski bagi Pembuat dan Penyebar Konten Pornografi, <,http://www.hukumonline.com/klinik/detail/lt540b73ac32706/sanksi-bagi-pembuat-dan-penyebar-konten-pornografi> [diakses pada 17 Mei 2017]

[3] Nasional.republika.co.id, Empat Pelaku Kasus Pornografi Anak di Medsos Diringkus Polisi http://nasional.republika.co.id/berita/nasional/daerah/17/03/15/omtfd6396-empat-pelaku-kasus-pornografi-anak-di-medsos-diringkus-polisi [diakses pada 18 Mei 2017]

[4] Pcbolong.blogspot.co.id, Web Crawler <http://pcbolong.blogspot.co.id/2011/04/web-crawler.html> [diakses pada 17 Mei 2017]

[5] Promptcloud.com, Few Pains of Web Crawling <https://www.promptcloud.com/few-pains-of-web-crawling/> [diakses pada 17 Mei 2017]

[6] Munir, Rinaldi. "Strategi Algoritmik", Program Studi Teknik Informatika, Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung, 2007.

[7] Mkyong.com, Jsoup Basic Web Crawler Example <https://www.mkyong.com/java/jsoup-basic-web-crawler-example/> [diakses pada 17 Mei 2017]

[8] Static.republika.co.id, Pornografi <http://static.republika.co.id/uploads/images/inpicture_slide/pornografi-_140701093539-739.jpg> [diakses pada 17 Mei 2017]

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 18 Mei 2017

Gianfranco Fertino Hwandiano
13515118